The secondary metabolite bioinformatics portal

Computational tools to facilitate synthetic biology of secondary metabolite production

Weber, Tilmann; Kim, Hyun Uk

Published in:
Synthetic and Systems Biotechnology

Link to article, DOI:
10.1016/j.synbio.2015.12.002

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Synthetic and Systems Biotechnology 1 (2016) 69–79

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: keaipublishing.com/synbio

The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production

Tilmann Weber [a,*], Hyun Uk Kim [a,b]

[a] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Alle 6, 2970 Hørsholm, Denmark
[b] BioInformatics Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 10 December 2015
Accepted 26 December 2015
Available online

Keywords:
Antibiotics
Biosynthesis
Bioinformatics
NRPS
PKS
Natural product

ABSTRACT

Natural products are among the most important sources of lead molecules for drug discovery. With the development of affordable whole-genome sequencing technologies and other 'omics tools, the field of natural products research is currently undergoing a shift in paradigms. While, for decades, mainly analytical and chemical methods gave access to this group of compounds, nowadays genomics-based methods offer complementary approaches to find, identify and characterize such molecules. This paradigm shift also resulted in a high demand for computational tools to assist researchers in their daily work. In this context, this review gives a summary of tools and databases that currently are available to mine, identify and characterize natural product biosynthesis pathways and their producers based on 'omics data. A web portal called Secondary Metabolite Bioinformatics Portal (SMBP at http://www.secondarymetabolites.org) is introduced to provide a one-stop catalog and links to these bioinformatics resources. In addition, an outlook is presented how the existing tools and those to be developed will influence synthetic biology approaches in the natural products field.

1. Introduction

Antimicrobial resistance is projected to be one of the major global challenges for maintaining our future health systems. According to the report commissioned by the Department of Health of the UK government, chaired by the economist Jim O'Neill, the global economic costs of antimicrobial resistance will result in more than 10 million annual deaths, leading to a loss of 2.0–3.5% of the world gross domestic product equivalent to 60–100 trillion USD by 2050 [e.g., references[1–3]]. While this report may predict a worst-case scenario, it is clear that the problem of antimicrobial resistance has to be urgently addressed globally. As there will be no simple single solution, efforts have to be undertaken in various fields, for example in optimizing hygiene, access to clear water, vaccinations, increased efforts to prevent infections, or reduced use of antibiotics families that are used in human medicine and feedstock.[4] Another important challenge will be to develop novel antimicrobial therapies and drugs.

Historically, natural products have been the major source of lead compounds for antimicrobial drugs,[5] but also are used in other application fields, such as anti-cancer drugs, insecticides, anthelmintics, painkillers, flavors, cosmeceuticals and crop protection. Nevertheless, most big pharma companies have severely reduced their research efforts on natural products during the last 20 years due to high rediscovery rates of known molecules and a lack of innovative screening approaches.[6] Therefore, it is surprising that still the majority of newly approved small-molecule drugs are natural products or their derivatives.[7]

With the broad availability of 'omics technologies, we currently experience a paradigm shift in natural product research; for decades, the only way to get access to new compounds was to cultivate antibiotics-producing microorganisms, mainly fungi and bacteria, under different growth conditions,[8] and then isolate and characterize the compounds with sophisticated analytical

Abbreviations: A, adenylation domain; BGC, biosynthetic gene cluster; C, condensation domain; GPR, gene-protein-reaction; HMM, hidden Markov model; LC, liquid chromatography; MS, mass spectrometry; NMR, nuclear magnetic resonance; NRP, non-ribosomally synthesized peptide; NRPS, non-ribosomal peptide synthetase; PCP, peptidyl carrier protein; PK, polyketide; PKS, polyketide synthase; RiPP, ribosomally and post-translationally modified peptide; SVM, support vector machine.

* Corresponding author. The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Alle 6, 2970 Hørsholm, Denmark. Tel.: +45 24 89 61 32; fax: +45 45 25 80 01.
E-mail address: tiwe@biosustain.dtu.dk (T. Weber).
Peer review under responsibility of KeAi Communications Co., Ltd.

http://dx.doi.org/10.1016/j.synbio.2015.12.002

chemistry. Nowadays, 'omics approaches offer complementary access to natural products; by identifying natural product/secondary metabolite biosynthetic gene clusters (BGCs), it is possible to assess the genetic potential of producer strains and to more effectively identify previously unknown metabolites. While this approach has led to some renaissance of natural product research in academia and industry, this information will also be the basis to rationally engineer molecules or develop "designer molecules" using synthetic biology approaches in the future.

When the first whole genome sequences of the model streptomycete *Streptomyces coelicolor* A3(2)[9] and the avermectin producer *Streptomyces avermitilis*[10,11] were determined, both strains were found to possess more secondary metabolite BGCs than an initial estimation made based on the number of their already known secondary metabolites. This is especially remarkable as both strains have served as model organisms and – in the case of *S. avermitilis* – industrial production strains for many years and thus have been studied by many researchers all over the world. With the rise of novel sequencing technologies and a growing number of microbial whole genome sequences, it became evident that a high number of BGCs is a common feature among various groups of bacteria, for example actinomycetes.[12]

Although the diversity of natural product chemical scaffolds is vast, the biosynthetic principles are highly conserved for many secondary metabolites. There is a set of enzyme families, which are often and very specifically associated with the biosynthesis of different classes of secondary metabolites. Thus, sequence information of these known gene families can be used to mine genomes for the presence of secondary metabolite biosynthetic pathways.

There are two principal strategies in the implementation of bioinformatic tools. Rule-based approaches can be used to identify gene clusters encoding known biosynthetic routes with high precision. In the first step of the mining process, these tools identify genes encoding conserved enzymes/protein domains that have associated roles in secondary metabolism, for example the "condensation (C)", "adenylation (A)" and "peptidyl carrier protein (PCP)" domains of non-ribosomal peptide synthetases (NRPSs). In the second step, predefined rules are used to associate the presence of such hits with defined classes of natural products. In the above example, a NRPS BGC can be simply and unambiguously identified if genes are present that code for at least one C-, A- and PCP domain. More complex rules may take into account whether specific genes are encoded in close proximity, for example type II polyketide BGCs can be detected using a rule that evaluates whether a ketosynthase α, a ketosynthase β/chain length factor and acyl-carrier protein are encoded by 3 individual genes in direct proximity. Such rule-based search strategies are, for example, implemented as one option in the pipeline antibiotics and Secondary Metabolite Analysis SHell (`antiSMASH`),[13–15] which, currently in its version 3, can detect 44 different classes of BGCs. Especially, clusters containing modular polyketide synthase (PKS) or NRPS genes can be easily detected by scanning the genome for genes that encode their characteristic enzyme domains, as also implemented in `NaPDoS`,[16] `NP.searcher`,[17] `GNP/PRISM`,[18] and `SMURF`.[19] All these approaches are very precise in detecting gene clusters of known families and classes of which rules can be defined. Based on the prerequisite to have defined rules, these algorithms cannot detect novel pathways that use a different biochemistry and enzymes. To avoid this limitation, also rule-independent methods, which are less biased, have been developed, for example implemented in `ClusterFinder`[20] and `EvoMining`[21] (see below for details on how they work). These tools use machine learning-based approaches or automated phylogenomics analyses to make their predictions. For fungi, algorithms that evaluate transcriptome data can also efficiently predict clusters of co-transcribed genes.[22]

As computational approaches to natural product discovery are rather a new and dynamic field, we intend to give an overview on existing computational tools and databases that help scientists solve the abovementioned tasks and develop perspectives on how these approaches will change the discovery of new natural products (Fig. 1).

2. Computational tools for natural product research

Recently, several reviews have been published, describing different strategies employed by the genome mining tools commonly used to detect secondary metabolite BGCs [e.g., references[23–26]]. In this review, we therefore give a summarizing, but comprehensive up-to-date overview on the tools and databases that are currently available for mining for BGCs, analyzing biosynthetic pathways, combining genomic and metabolomic data, and generating genome-scale metabolic models of the secondary metabolite producers (Tables 1 and 2). More importantly, this overview information is coherently provided through the newly established Secondary Metabolite Bioinformatics Portal (SMBP) along with links to references and websites of the tools and databases. We also discuss perspectives on further development of the field.

2.1. Manual genome mining

Before automated tools (see below) became available, genome mining approaches have been undertaken by "manually" identifying key biosynthetic enzymes in genome data. For this, either amino acid sequences of characterized proteins of interest were used as queries for `BLAST` or `PSI-BLAST`,[75] or – if alignments of a family of query sequences were available – these were used to generate profile Hidden Markov Models (HMMs) which served as queries using the software `HMMer`.[76] Gene clusters were then identified by analyzing the genes encoded up- and downstream of the hit sequence. While this approach has been superseded by automatic tools for most of the commonly observed gene cluster types, it is still highly relevant for identifying gene clusters which are not covered by the rulesets of the common tools and where prototypes have just been discovered and described. The manual genome mining can be further improved with tools like `MultiGeneBlast`,[77] which allow a BLAST-based analyses of whole operons or gene clusters.

2.2. Tools for identification of BGCs

Identifying BGCs with `BLAST` and `HMMer` works very well with low false positive rates for many different classes of secondary metabolites, for example polyketides (PKs) synthesized by type I or type II PKS, ribosomally and post-translationally modified peptides (RiPPs), or NRPs. Therefore, a number of tools have been developed that use rule-based approaches, i.e., the specific search for distinct enzymes or enzymatic domains (Fig. 1).

`BAGEL`[28–30] is a web-based comprehensive mining suite to identify and characterize RiPPs in microbial genomes. `BAGEL` provides an annotation-independent identification of the genes encoding precursor peptides, classification of the RiPP types as well as a database of known RiPPs. Especially, in the field of identification of the BGCs of type I PKS, NRPS and hybrid PKS/NRPS, a wide variety of tools exist. `ClustScan`[39] is a Java-based desktop application that offers mining for PKS and NRPS gene clusters in a convenient graphical user interface. `ClustScan` was used to compile and analyze the data contained in the `ClustScan` database (see below). `NP.searcher`[17] is a web-based software program with an emphasis on structure prediction of the putative peptide or polyketide metabolites. `NaPDoS`[16] uses `BLAST` and `HMMer` to identify ketosynthase domain (in PKS) and condensation domain (in NRPS) encoding genes in genomic and metagenomic datasets and provides a detailed phylogenetic analysis of these domains which are then classified into functional categories. `GNP/Genome search`[35,69,78] and `GNP/PRISM`[18] are web-based tools to mine for and analyze PKS and NRPS pathways,

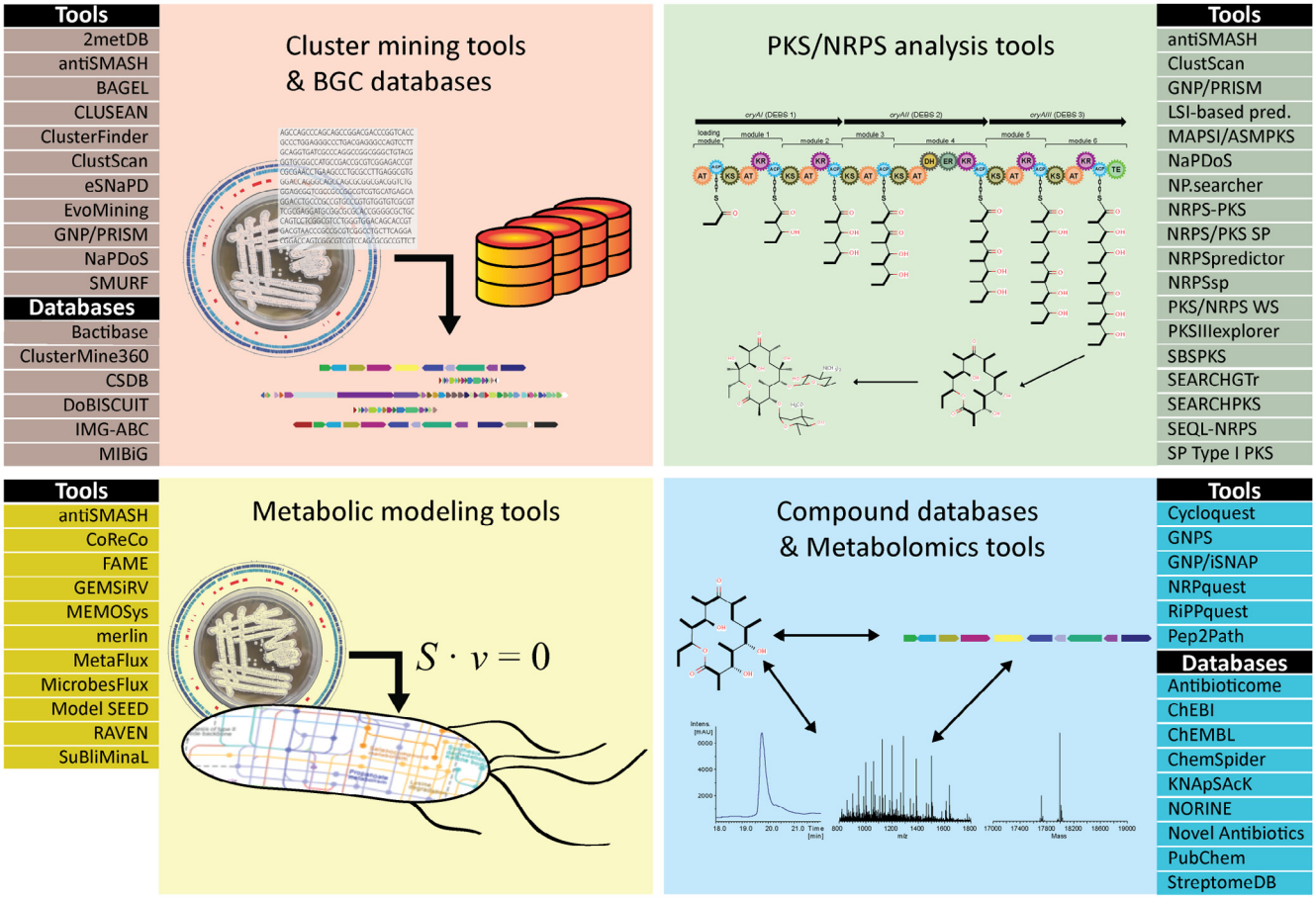

Fig. 1. Overview of the most commonly used and freely accessible tools specialized for the analysis of secondary metabolites and their pathways.

including identification of similar known pathways, the latter with an emphasis on the prediction of putative products. They are closely interconnected with the metabolomics platform iSNAP, which uses information on predicted products to identify corresponding peaks in liquid chromatography/tandem mass spectrometry (LC-MS/MS) data (see paragraph 2.6). The Secondary Metabolite Unknown Region Finder SMURF[19] can detect fungal PKS, NRPS and terpenoid gene clusters involving a dimethylallyltryptophan synthase type prenyltransferases. With pipelines such as CLUster SEquence ANalyzer (CLUSEAN),[31] there are also tools available that can automate the analysis of larger datasets using scripts instead of interactive web pages.

While the tools mentioned above are specialized in detecting and analyzing specific classes of secondary metabolites, antiSMASH[13–15] provides detection rules for 44 different classes and subclasses of secondary metabolites. In addition to the identification of gene clusters, antiSMASH also provides detailed annotation of the domain structures of modular PKS and NRPS, analysis of lanthipeptide pathways,[79] substrate predictions, genome-scale metabolic modeling and comparative genomics tools to identify conserved subclusters biosynthesizing building-blocks, similar gene clusters in other sequenced genomes and the Minimum Information about a Biosynthetic Gene cluster (MIBiG)-standard[56] dataset. With this functionality, antiSMASH currently is the most comprehensive software for mining microbial genomes for BGCs. In the future, it is planned to extend antiSMASH as a generic platform integrating various tools such as CRISPy-web, a web-based tool to design guide RNAs (sgRNAs) for CRISPR applications (Blin et al. in this issue).

All rule-based BGC-mining approaches can precisely identify BGCs of known biosynthetic types, but fail to identify pathways, which use non-homologous enzymes or enzymes with biochemistry that is presently unknown. However, there are some alternative approaches that try to identify BGCs independent of pre-defined rulesets. The software ClusterFinder,[20] which also is implemented as an alternative cluster detection algorithm in antiSMASH, uses a HMM-based approach to detect chromosomal regions in genomes that aggregate protein domains associated with secondary metabolite biosynthetic pathways. The EvoMining approach[21] identifies gene clusters based on the observation that many BGCs encode isoenzymes closely related to primary metabolism, but displaying a different phylogeny. By scanning the genomes for the occurrence of such enzymes, it is possible to detect secondary metabolite BGCs without respect to their conserved enzymology.

2.3. Tools for analyzing specific enzymes

In addition to the general genome mining tools mentioned above, a whole set of tools was developed specifically to provide automated specificity prediction for NRPS A-domains and to detect the enzymatic domains in multi-modular PKS and NRPS, such as SEARCHPKS[42] or NRPS-PKS/SBSPKS.[40,41] One of the hallmarks of computational analysis of secondary metabolite biosynthetic pathways was the deciphering of the NRPS A-domain specificity conferring code by Stachelhaus et al.[80] and Challis et al.,[81] who found out that conserved amino acids near the active site of NRPS A-domains can be used to map the substrate specificity of these enzymes, which is an important prerequisite for the computational

Table 1
Comprehensive collection of freely accessible software programs and databases dedicated to natural product research. Only software programs and databases properly functioning as of December 2015 are listed in this table. A more comprehensive list can be found at the SMBP (http://www.secondarymetabolites.org).

Software program or database	URL	Reference	Last publication or documented update	Main content and/or function
Tools for mining of secondary metabolite gene clusters				
[R]: rule-based, [N]: non-rule based algorithms used to detect the BGCs				
2metDB[R]	http://secmetdb.sourceforge.net/	27	2013	Standalone (Mac) tool to mine PKS/NRPS gene clusters
antiSMASH[R/N]	http://antismash.secondarymetabolites.org	13–15	2015	Web application and standalone tool (LINUX, MacOS and MS Windows) to mine and analyze BGCs; includes comparative genomics tools and a homology-based metabolic modeling pipeline
BAGEL[R]	http://bagel2.molgenrug.nl/	28–30	2013	Web application to mine and analyze RiPPs
CLUSEAN[R]	https://bitbucket.org/tilmweber/clusean	31	2013	Standalone (LINUX and MacOS) tool to mine and analyze BGCs, mainly PKS/NRPS
ClusterFinder[N]	https://github.com/petercim/ClusterFinder	20	2014	Standalone tool (LINUX and MacOS) to identify BGCs with an non-rule based approach
eSNaPD[R]	http://esnapd2.rockefeller.edu/	32–34	2014	Web application to mine metagenomic datasets for BGCs
EvoMining[N]	http://148.247.230.39/newevomining/new/evomining_web/index.html	21	2015	Web application for phylogenomic approach of cluster identification
GNP/Genome Search[R]	http://magarveylab.ca/gnp/#!/genome	35	2015	Web application to mine and analyze BGCs, mainly PKS/NRPS
GNP/PRISM[R]	http://magarveylab.ca/prism	18	2015	Web application to mine and analyze BGCs, mainly PKS/NRPS, including glycosylations and structure prediction
MIDDAS-M[N]	http://133.242.13.217/MIDDAS-M/	36	2013	Web application to use transcriptome data to identify BGC coordinates in fungal genomes
MIPS-CG[N]	http://www.fung-metb.net/	37,38	2015	Web application to identify BGC coordinates in fungal genomes without transcriptome data
NaPDoS[R]	http://napdos.ucsd.edu/	16	2012	Web application offering phylogenomic analysis of PKS-KS and NRPS-C domains
SMURF[R]	http://jcvi.org/smurf/index.php	19	2010	Web application to mine PKS/NRPS/terpenoid gene clusters in fungal genome
Software for the analysis of type I PKS and NRPS pathways				
ClustScan Professional	http://bioserv.pbf.hr/cms/index.php?page=clustscan	39	2008	Java-based standalone tool to mine for PKS/NRPS BGCs
NP.searcher	http://dna.sherman.lsi.umich.edu/	17	2009	Web application/standalone tool (LINUX) to mine for PKS/NRPS BGCs
NRPS-PKS/SBSPKS	http://www.nii.ac.in/~pksdb/sbspks/master.html	40,41	2010	Web application to mine for PKS BGCs
SEARCHPKS	http://linux1.nii.res.in/~pksdb/DBASE/pagesearchpks.html	42	2003	Web application to mine for PKS BGCs
Software for predicting substrate specificities				
LSI-based A-domain function predictor	http://bioserv7.bioinfo.pbf.hr/LSIpredictor/AdomainPrediction.jsp	43	2014	Web application to predict A-domain specificities
NRPS/PKS substrate predictor	http://www.cmbi.ru.nl/NRPS-PKS-substrate-predictor/	44	2013	Web application to predict A-domain/AT-domain specificities
NRPSpredictor/NRPSpredictor2	http://nrps.informatik.uni-tuebingen.de	45,46	2011	Web application/standalone tool (LINUX, MS Windows, MacOS) to predict A-domain specificities
NRPSsp	http://www.nrpssp.com/	47	2012	Web application to predict A-domain specificities
PKS/NRPS Web Server/Predictive Blast Server	http://nrps.igs.umaryland.edu/nrps/	27	2009	Web application to determine domain organization and A-domain specificities
SEARCHGTr	http://linux1.nii.res.in/~pankaj/gt/gt_DB/html_files/searchgtr.html	48	2005	Web application to predict glycosyltransferase specificities
SEQL-NRPS	http://services.birc.au.dk/seql-nrps/	49	2015	Web application to predict A-domain specificities

(*continued on next page*)

Table 1 (*continued*)

Software program or database	URL	Reference	Last publication or documented update	Main content and/or function
Databases focusing on gene clusters				
Bactibase	http://bactibase.pfba-lab-tun.org	50,51	2011	Web accessible database of bacteriocins
ClusterMine360	http://www.clustermine360.ca/	52	2013	Web accessible database of BGCs
ClustScan Database	http://csdb.bioserv.pbf.hr/csdb/ClustScanWeb.html	53	2013	Web accessible database of PKS/NRPS BGCs
DoBISCUIT	http://www.bio.nite.go.jp/pks/	54	2015	Web accessible database of PKS/NRPS BGCs
IMG-ABC	http://img.jgi.doe.gov/abc	55	2015	Web accessible database of BGCs, tightly integrated into JGI's IMG platform
MIBiG	http://mibig.secondarymetabolites.org	56	2015	Web accessible repository of BGCs
Recombinant ClustScan Database	http://csdb.bioserv.pbf.hr/csdb/RCSDB.html	57	2013	Database of *in silico* recombined BGCs
Databases focusing on bioactive compounds				
Antibioticome	http://magarveylab.ca/antibioticome	Unpublished	2015	Web accessible database on compounds, compound families and modes of action
ChEBI	https://www.ebi.ac.uk/chebi/	58	2015	Web accessible database and ontology on compounds focused on small molecules
ChEMBL	https://www.ebi.ac.uk/chembl/	59	2015	Web accessible database on bioactive compounds with drug-like properties
ChemSpider	http://www.chemspider.com/	60	2015	Web accessible database on structures and properties of over 35 million structures
KNApSAcK database	http://kanaya.aist-nara.ac.jp/KNApSAcK/	61,62	2015	Web accessible database on compounds; standalone version of KNApSAcK metabolite database available
NORINE	http://bioinfo.lifl.fr/norine	63,64	2015	Web accessible database on NRPs
Novel Antibiotics Database	http://www.antibiotics.or.jp/journal/database/database-top.htm	Unpublished	2008	Web accessible database on compounds
PubChem	http://pubchem.ncbi.nlm.nih.gov/	65	2015	Web accessible database on compounds and bioactivities; source data available for download
StreptomeDB	http://www.pharmaceutical-bioinformatics.de/streptomedb	66,67	2015	Web accessible database on compounds produced by streptomycetes; download of compounds and metadata in SD format.
Metabolomics tools				
Cycloquest	http://cyclo.ucsd.edu	68	2011	Web application to correlate tandem MS data of cyclopeptides with gene clusters
GNPS	http://gnps.ucsd.edu/	unpublished	2015	Generic metabolomics portal to analyze MS/MS data (dereplication and molecular networking)
GNP/iSNAP	http://magarveylab.ca/gnp/	35,69–71	2015	Web application to automatically identify metabolites in MS/MS data based on genomic data
NRPquest	http://cyclo.ucsd.edu	72	2014	Web application to correlate NRP tandem data with gene clusters
Pep2Path	http://pep2path.sourceforge.net	73	2014	Standalone application to correlate peptide sequence tags with NRP and RiPP BGCs
RiPPquest	http://cyclo.ucsd.edu	74	2014	Web application to correlate RIPP tandem data with gene clusters

Table 2
High-throughput metabolic modeling tools that can facilitate engineering of actinomycetes for secondary metabolite production. Tools are shown in the order of the year they appeared.

Software program	URL	Reference	Year of publication	Main content and/or function
Model SEED	http://seed-viewer.theseed.org/seedviewer.cgi?page=ModelView	99	2010	First online high-throughput metabolic modeling tool
MEMOSys	https://memosys.i-med.ac.at/MEMOSys/home.seam	100	2011	Allows management, storage, and development of metabolic models
SuBliMinaL Toolbox	http://www.mcisb.org/resources/subliminal/	101	2011	Has strengths in managing chemical information for metabolites in a metabolic model
FAME	http://f-a-m-e.fame-vu.vm.surfsara.nl/ajax/page1.php	102	2012	Allows streamlined analysis of a newly built metabolic model using various simulation methods
GEMSiRV	http://sb.nhri.org.tw/GEMSiRV/en/GEMSiRV	103	2012	Allows metabolic model reconstruction, simulation and visualization
MetaFlux in Pathway Tools	http://bioinformatics.ai.sri.com/ptools/	104	2012	Provides strong supports for predicting, modeling, curating and visualizing metabolic pathways
MicrobesFlux	http://www.microbesflux.org/	105	2012	Allows both flux balance analysis (FBA) and dynamic FBA of a newly generated metabolic model
RAVEN Toolbox	http://biomet-toolbox.org/index.php?page=downtools-raven	106	2013	Allows metabolic model reconstruction, simulation and visualization in MATLAB environment
CoReCo	https://github.com/esaskar/CoReCo	107	2014	Useful for modeling metabolisms of multiple related species
merlin	http://www.merlin-sysbio.org/	108	2015	Most recently released metabolic modeling program with comprehensive genome annotation functionalities necessary for model generation
antiSMASH	http://www.secondarymetabolites.org	13	2015	Provides comprehensive genome mining platform for BGCs; currently the only platform offering automated modeling including secondary metabolite specific reactions

prediction of the biosynthetic products. The PKS/NRPS Web Server, Predictive Blast Server, and 2metDB[27] deliver predictions based on BLAST analyses against the signatures determined by Challis et al.[81] Later tools introduced the use of profile HMMs, for example an algorithm by Minowa et al.,[82] NRPSsp,[47] NRPS/PKS substrate predictor,[44] machine learning-based on transductive Support Vector Machines (SVMs), as for example implemented in NRPSpredictor,[45,46] Latent Semantic Indexing, which is used by the LSI-based A-domain predictor[43] or the Sequence Learner algorithm, which is used in SEQL-NRPS.[49] There have also been first successful reports on using structural bioinformatics involving both crystal structure or homology models and docking analyses with putative substrates, which contributed to predicting substrate specificities of A-domains.[83] However, this approach is currently very compute-intensive, and no automated tools have been reported so far. For other enzymes involved in secondary metabolite biosynthesis, only few tools are available. PKSIIIexplorer[84] uses transductive SVMs to classify type III PKSs. SEARCHGTr[48] currently is the only tool that offers prediction of glycosyltransferase specificities.

2.4. Databases focusing on biosynthesis genes and their clusters

All the tools mentioned in the previous section can be used to identify or analyze secondary metabolite BGCs or specific enzymes of the pathways in the user-submitted gene cluster/genome data. To allow cross-species comparison, several databases have been developed focusing on different aspects of secondary metabolism. The ClustScan database,[53] DoBISCUIT,[54] and ClusterMine360[52] provide collections of a limited set of mostly hand-curated PKS and NRPS gene clusters. The recombinant ClustScan database r-CSDB[57] in addition contains more than 20,000 *in silico* recombined sequences that are expected to produce novel molecules. Recently, a standard on MIBiG has been developed.[56] In the course of this project, a MIBiG repository was generated, containing more than 1000 characterized BGCs; more than 400 of them were manually annotated and curated by the original researchers carrying out the experimental characterizations. In addition to these databases, data collections were also established based on large-scale sequencing efforts. The Integrated Microbial Genomes: Atlas of Biosynthetic Gene Clusters (IMG-ABC)[55] is a huge data collection based on manually curated BGCs, but also includes automatically mined BGCs of public genome data and genomes that were sequenced at the US Department of Energy Joint Genome Institute (JGI). Currently, IMG-ABC is the largest collection of BGCs data.

So far, the genome data used for genome mining of whole biosynthetic pathways almost exclusively originated from cultivable organisms. Considering the fact that only a little percentage of environmental bacteria can be grown in culture, the unculturable microorganisms remain a huge and currently under-exploited resource. The environmental Surveyor of NAtural Product Diversity (eSNAPD)[32,33,85] is a system to map amplicon datasets to known BGCs. As eSNAPD can also use location metadata, the data can be analyzed based not only on the sequences but also on location information about the sampling sites.

2.5. Databases focusing on compounds

In addition to general public molecule databases, such as PubChem,[65] ChEMBL,[86,87] and ChEBI,[88,89] which contain information on a humongous volume of chemical compounds including secondary metabolites, commercial natural product compound databases are available, including antiBASE (Wiley-VCH, Weinheim, Germany), and the Dictionary of Natural Products (Taylor and Francis Group LLC, USA). Recently, several freely accessible or openly licensed databases have also been developed. The KNApSAcK[61,62] website offers information on various secondary metabolites with respect to their basic chemical properties and bioactivities. Although the KNApSAcK system is mostly focused on plant metabolites, it also contains information on microbial bioactive compounds. A component of the KNApSAcK system dealing with metabolites can also be downloaded and used as a standalone Java-based tool. StreptomeDB[66,67] is a database focusing on secondary metabolites isolated from streptomycetes. Bactibase[50,51] is focused on ribosomally synthesized antimicrobial peptides, while NORINE[63,64] is a hand-curated database of NRPs and their activities.

2.6. Metabolomics tools for natural product identification

LC-MS and nuclear magnetic resonance (NMR)-based metabolomics approaches gain increasing importance in natural product studies [for reviews, see references[90,91]]. While some of the tools or databases on natural product compounds and their BGCs already have histories of more than ten years, first computational approaches have been published only very recently that use cheminformatic approaches to automatically classify and map metabolomics (i.e., MS and MS/MS data) to natural product families and corresponding biosynthetic pathways. This has been especially successful for identifying peptides (RiPPs and NRPs) in the mass spectra of complex samples. Software programs for these approaches include Pep2Path,[73] RiPPquest,[74] NRPquest,[72] and Cycloquest.[68] The GNP/iSNAP (From Genes to Natural Products) – web application provides a user-friendly interface to carry out analyses of MS/MS data of NRP producing strains.[35,69,70] Signals corresponding to NRPs or NRP-analogs are detected by comparison to databases containing computationally generated fragments of known secondary metabolites (e.g., those extracted from NORINE[63] or PubChem[65]). Recently, iSNAP has also been extended to identify PK compounds and analogs of known molecules.[70]

The Global Natural Products Social Molecular Networking system (GNPS) provides workflows for automated spectra deconvolution, molecular networking to identify compound families and dereplication against a database of known molecules (unpublished). In addition to the analysis function, GNPS has a social network component that allows users to share their mass spectrometry datasets (including continuous identification by re-analyzing the deposited datasets against updated spectra libraries) or datasets of reference compounds.

2.7. High-throughput metabolic modeling tools

The availability of genomic information allows generation of genome-scale metabolic models, which have now become one of standard tools in systems biology and metabolic engineering communities. This technology enables linking between genotype, including BGCs of secondary metabolites, and metabolic phenotype of secondary metabolite producing microorganisms. A genome-scale metabolic model is a type of mathematical model that is based on mass balances of all the metabolites known/predicted to be present in an organism of interest and is represented in a large-scale stoichiometric matrix that can be simulated with various numerical optimization tools.[92] One of the unique features of genome-scale metabolic model is description of gene-protein-reaction (GPR) associations in a Boolean format; the GPR associations logically connect genomic information with the organism's metabolism, and hence enable prediction of various metabolic phenotypes using gene-level information. In the field of secondary metabolites, genome-scale metabolic models have largely contributed to studies on (i) predicting intracellular flux distributions of actinomycetes under specific environmental/genetic conditions[93,94] and (ii) gene manipulation targets for overproduction of target secondary metabolites.[95,96]

Although development of a genome-scale metabolic model is a laborious and time-consuming procedure, involving a total of 96 steps in a protocol,[97] a large fraction of the procedure can now be automated. Such high-throughput metabolic modeling tools allow streamlined system-wide metabolic studies for newly sequenced genomes of actinomycetes and other secondary metabolite producers whose number keeps growing due to increased attentions on novel antibiotics discovery. Among currently available high-throughput metabolic modeling tools, to our knowledge, only Model SEED has been deployed to reconstruct multiple actinomycete species in a high-throughput manner for large-scale metabolic studies.[98] Currently available high-throughput modeling tools are summarized in Table 2. For a detailed comparison of high-throughput metabolic modeling tools, see Hamilton and Reed,[109] and Dias et al.[108] Finally, a challenge for modeling secondary metabolite producers is that all the available metabolic modeling tools do not consider secondary metabolite biosynthesizing reactions and their relevant precursors, and the fact that most secondary metabolites are biosynthesized in stationary phase and not in the exponential growth phase, which stands against the pseudo-steady state assumption of this modeling approach. These special circumstances will therefore require additional efforts in optimizing the metabolic models.

3. The secondary metabolite bioinformatics portal

The field of secondary metabolite bioinformatics is drastically changing with new tools being released and old services discontinued. We therefore started the web-portal SMBP as a one-stop access point containing a manually curated collection of all the relevant tools and databases for 'omics-based secondary metabolism research, including short descriptions of the tools, literature references and links to the web sites and/or download pages (Fig. 2). Currently, the tools and databases are assigned to one (or more) categories of contents/functionalities covering secondary metabolite compounds, genome mining, PKS/NRPS analysis, specificity predictors, metabolomics analysis, metabolic modeling and generic tools. A full text search engine provides easy access to the relevant information. The SMBP is openly available at http://www.secondarymetabolites.org, and the Markdown source code for the portal is available at https://bitbucket.org/secmetbioinf/portal.

4. Future challenges

Despite significant advances on computational approaches to identify and characterize BGCs, there still exist several challenges that have to be addressed in the near future.

Even for the well-studied secondary metabolite classes such as PK or NRP pathways, prediction of the core scaffold structure of a compound is incomplete because the biochemical knowledge on these systems is not yet implemented in the software (relatively easy to fix in this case) or the relevant biochemical knowledge is not sufficiently available to be the basis for the implementation of novel computational algorithms (more difficult to overcome than the former case). In particular, for machine learning-based approaches, the availability of medium- to large-scale biochemical data required to train good models is very limiting in many cases.

Another unsolved problem is currently inaccurate prediction of gene cluster borders. The most widely used genome mining software antiSMASH simply assigns n kb upstream or downstream of the core biosynthetic genes to the cluster (for example, $n = 20$ kb for PKS and NRPS clusters, and $n = 10$ kb for lanthipeptides). SMURF, which addresses fungal PK, NRP and terpenoid metabolites, uses a different approach; a statistical analysis of 22 clusters of the model strain *Aspergillus fumigatus* led to the identification of a total of 27 protein domains, which commonly co-occur with the PK, NRP and terpenoid biosynthetic genes. The occurrence of these domains in genes flanking the core biosynthetic genes, together with the intergenic distance, is then considered to calculate the cluster borders.[19] Another promising approach to predict BGC borders is to use comparative genomics data; genes within a putatively identified BGC, which are conserved among other producers of similar compounds, are likely to belong to the BGCs, whereas genes not belonging to the cluster are more divergent. An algorithm implementing this strategy for filamentous fungi (MIPS-CG) has been described by Takeda et al.[37] For fungal BGCs, it has further been demonstrated that – in addition to the mining and analysis methods described above – transcriptome data can provide valuable

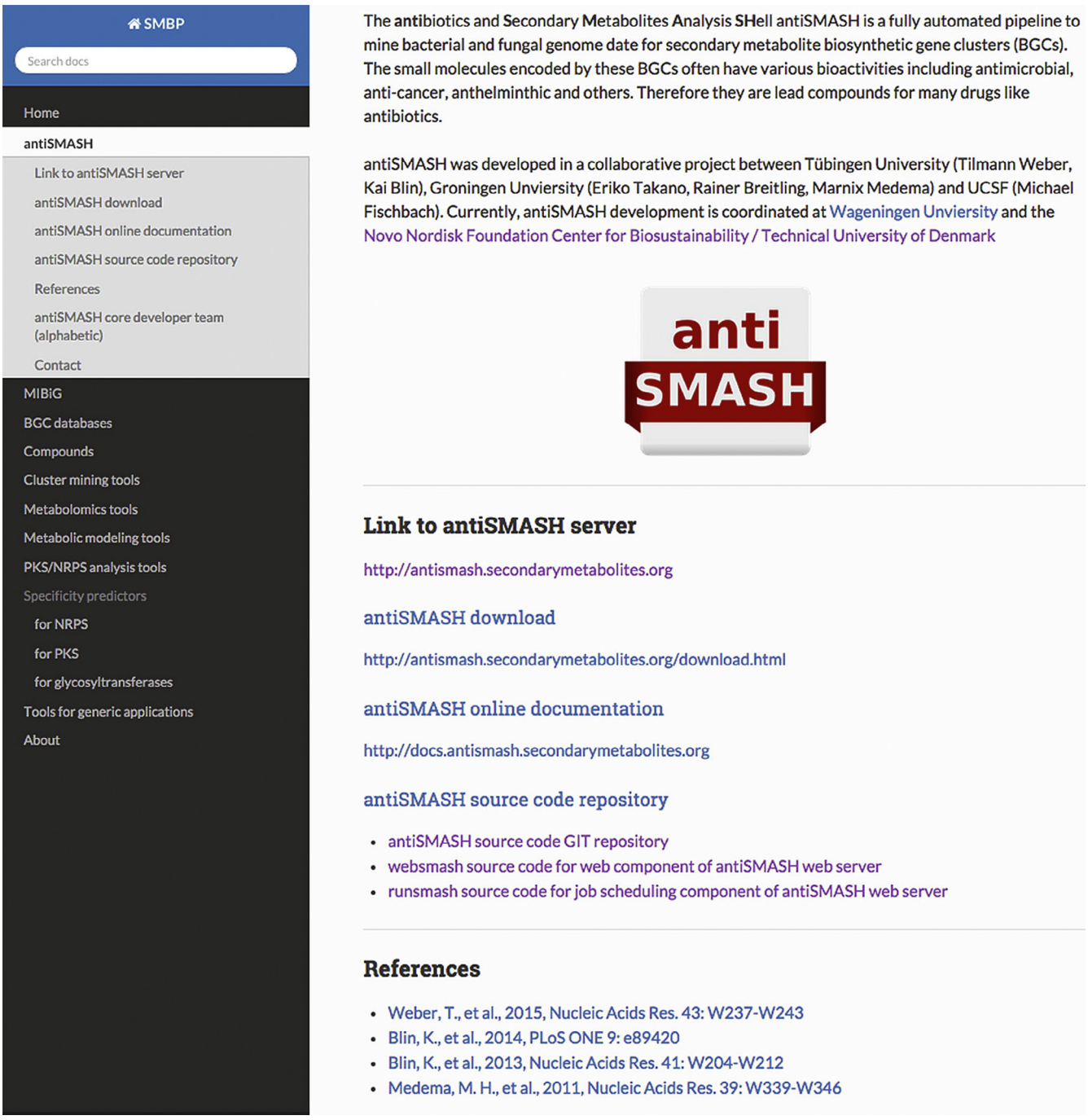

Fig. 2. A screenshot of the antiSMASH page in the Secondary Metabolite Bioinformatics Portal at http://www.secondarymetabolites.org.

information on the borders of the BGCs.[22,36] For prokaryotes, to our best knowledge, no such observations have been reported so far.

Analyses involving the integration of different "kinds" of data (e.g., genome with transcriptome or metabolome data) generally suffer from a very poor integration of different functionalities available across the tools and the requirement of specific input and output formats; all these barriers make using relevant software programs difficult for researchers not familiar with bioinformatics. In fact, this is a chronic problem in bioinformatics and systems biology in general. Advances in integrating heterogeneous 'omics data would offer new dereplication opportunities to identify already known metabolites at a very early stage of the metabolite discovery process. In relation to this, proteome data can deliver important information on secondary metabolite biosynthesis when they are correlated to metabolome data (e.g., obtained by LC-MS) and bioactivity profiles. Using a set of different growth conditions, which leads to the differential expression of BGCs and thus different bioactivity profiles, Gubbens et al.[110] were able to correlate the expression levels of biosynthetic enzymes with the occurrence of secondary metabolites. Using this approach, it was possible to identify juglomycin C and the corresponding gene cluster in *Streptomyces* sp. MBT70.[110] Furthermore, the power of combining large-scale genome and metabolome data was explored along with computational approaches to identify novel secondary metabolites.[111] Doroghazi et al. identified 11,422 PKS-, NRPS-, NRPS-independent siderophores, lanthipeptides and thiazole-oxazole modified microcin gene

clusters in 830 genome sequences of actinomycetes. The gene cluster sequences were then clustered based on a combination of different distance metrics, resulting in 4122 gene cluster families. For a subset of 178 analyzed strains, this network was then automatically correlated with high-resolution mass spectrometric data of known compounds leading to the automatic identification of 110 molecules and 27 molecule families. Thus, for some of these molecule families, previously unidentified gene clusters could be automatically related to the produced metabolite. Taken together, as demonstrated in the studies discussed above, it is highly desirable to interconnect the existing tools and data, and automate the analysis workflows for streamlined characterization of genomes and their resulting secondary metabolites. Current bottlenecks in such integrative approaches can be relieved by standardizing APIs and data structures for programmatic access of the different tools.

5. Implications for synthetic biology applications in natural product studies

While the availability of computational tools provides new possibilities for identifying and characterizing novel secondary metabolites, such tools are also essential for the development of synthetic biology strategies, which aim at the efficient production of rationally designed molecules.[112] While there exist several generic synthetic biology tools to predict, prioritize, model, select and implement pathways, as reviewed in reference[113], only few reports exist on their use to engineer natural product biosynthetic pathways.

Especially, engineering PKS and NRPS megasynthases will need further emphasis; from a formal perspective, these modular enzymes are excellent candidates for synthetic biology approaches because they display a modular organization and a well-defined split-up of “enzymatic tasks” and tempt to easy plug-and-play approaches. Although there are many successful module and/or domain replacements reported during the last 15 years that led to rationally [e.g., references[114–118]] or combinatorially [e.g., references[119]] engineered products, the failure rates are still high and the yields obtained with the engineered assembly lines usually decrease severely. The main reason for this is likely that for designing the modified enzymes, mostly sequence divergence at the linker regions between the enzymatic domains or even trial-and-error approaches might have caused the suboptimal performance of the engineered assembly lines (i.e., inactivity or drastically decreased yields) as they interfered with the 3D structure and the intra- and intermolecular protein–protein interactions within the highly complex megaenzymes. Because structural data of not only separate enzymatic domains but also complete modules for both NRPS[120,121] and type I PKS[122,123] recently became available, they now offer the molecular background to overcome current challenges in engineering the PKS or NRPS assembly lines. In the same line, biochemical studies have been carried out, which specifically address how different domains interact with one another within the PKS or NRPS assembly lines and may help better understanding of the molecular mechanisms within the assembly lines [e.g., references[124,125]]; this knowledge has yet to be integrated into synthetic biology design software. Certainly, these approaches will be supported by the availability of heterologous expression and genome engineering tools like CRISPR, which recently also became available for secondary metabolite producers.[126–129] These technologies will drastically reduce the efforts to generate the required recombinant strains and thus allow the high-throughput generation of many variants.

6. Conclusions

Genome mining and other ‘omics-based approaches to identify and characterize secondary metabolites and their producers have become essential technologies complementing the classical approaches of natural product discovery. This trend is manifested by an increasing number of new and improved bio- and cheminformatic tools and databases bridging computational biology and wet-lab work in the field. Because of the ever-growing number of computational tools and databases dedicated to secondary metabolites, we herein release the SMBP (http://www.secondarymetabolites.org) where researchers in the field can explore diverse tools and databases in one stop. The SMBP is expected to enable users to compare tools for their utilities and make further contributions to the field of secondary metabolites.

Acknowledgments

The work of the authors is supported by a grant of the Novo Nordisk Foundation, Denmark.

References

1. Taylor J, Hafner M, Yerushalmi E, Smith R, Bellasio J, Vardavas R, et al. Estimating the economic costs of antimicrobial resistance. Models and results. California: RAND Corporation SM and Cambridge, UK: The Wellcome Trust; 2014.
2. KPMG LLP UK. Report on the global economic impact of anti-microbial resistance, commissioned by the Wellcome Trust; 2014.
3. The Review on Antimicrobial Resistance. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. London: The Review on Antimicrobial Resistance; 2014. <http://amr-review.org/sites/default/files/SECURING%20NEW%20DRUGS%20FOR%20FUTURE%20GENERATIONS%20FINAL%20WEB_0.pdf>.
4. Bush K, Courvalin P, Dantas G, Davies J, Eisenstein B, Huovinen P, et al. Tackling antibiotic resistance. *Nat Rev Microbiol* 2011;**9**:894–6.
5. Cragg GM, Newman DJ. Natural products: a continuing source of novel drug leads. *Biochim Biophys Acta* 1830;**2013**:3670–95.
6. Fischbach MA, Walsh CT. Antibiotics for emerging pathogens. *Science* 2009;**325**:1089–93.
7. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 2012;**75**:311–35.
8. Bode HB, Bethe B, Hofs R, Zeeck A. Big effects from small changes: possible ways to explore nature’s chemical diversity. *Chembiochem* 2002;**3**:619–27.
9. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 2002;**417**:141–7.
10. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, et al. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* 2003;**21**:526–31.
11. Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* 2001;**98**:12215–20.
12. Weber T, Charusanti P, Musiol-Kroll EM, Jiang X, Tong Y, Kim HU, et al. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. *Trends Biotechnol* 2015;**33**:15–26.
13. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0 – a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 2015;**43**:W237–43.
14. Blin K, Medema MH, Kazempour D, Fischbach M, Breitling R, Takano E, et al. antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res* 2013;**41**:W204–12.
15. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011;**39**:W339–46.
16. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE* 2012;**7**:e34064.
17. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. *BMC Bioinformatics* 2009;**10**:185.
18. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res* 2015;doi:10.1093/nar/gkv1012.
19. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* 2010;**47**:736–41.
20. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 2014;**158**:412–21.
21. Cruz-Morales P, Martínez-Guerrero CE, Morales-Escalante MA, Yáñez-Guerra LA, Kopp JF, Feldmann J, et al. Recapitulation of the evolution of biosynthetic

gene clusters reveals hidden chemical diversity on bacterial genomes. *bioRxiv* 2015;doi:10.1101/020503.
22. Andersen MR, Nielsen JB, Klitgaard A, Petersen LM, Zachariasen M, Hansen TJ, et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc Natl Acad Sci U S A* 2013;**110**:E99–107.
23. Bachmann BO, Van Lanen SG, Baltz RH. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol* 2014;**41**:175–84.
24. Weber T. *In silico* tools for the analysis of antibiotic biosynthetic pathways. *Int J Med Microbiol* 2014;**304**:230–5.
25. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol* 2015;**11**:639–48.
26. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. *J Ind Microbiol Biotechnol* 2014;**41**:443–50.
27. Bachmann BO, Ravel J. Chapter 8. Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 2009;**458**:181–217.
28. van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. *Nucleic Acids Res* 2013;**41**:W448–53.
29. de Jong A, van Heel AJ, Kok J, Kuipers OP. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res* 2010;**38**:W647–51.
30. de Jong A, van Hijum SA, Bijlsma JJ, Kok J, Kuipers OP. BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Res* 2006;**34**:W273–9.
31. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol* 2009;**140**:13–7.
32. Charlop-Powers Z, Owen JG, Reddy BV, Ternei MA, Guimaraes DO, de Frias UA, et al. Global biogeographic sampling of bacterial secondary metabolism. *Elife* 2015;**4**:e05048.
33. Owen JG, Reddy BV, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH, et al. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A* 2013;**110**:11797–802.
34. Reddy BV, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z, Brady SF. Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl Environ Microbiol* 2012;**78**:3744–52.
35. Johnston CW, Skinnider MA, Wyatt MA, Li X, Ranieri MR, Yang L, et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat Commun* 2015;**6**:8421.
36. Umemura M, Koike H, Nagano N, Ishii T, Kawano J, Yamane N, et al. MIDDAS-M: Motif-Independent *De Novo* Detection of Secondary Metabolite gene clusters through the integration of genome sequencing and transcriptome data. *PLoS ONE* 2013;**8**:e84028.
37. Takeda I, Umemura M, Koike H, Asai K, Machida M. Motif-independent prediction of a secondary metabolism gene cluster using comparative genomics: application to sequenced genomes of *Aspergillus* and ten other filamentous fungal species. *DNA Res* 2014;**21**:447–57.
38. Umemura M, Koike H, Machida M. Motif-independent de novo detection of secondary metabolite gene clusters-toward identification from filamentous fungi. *Front Microbiol* 2015;**6**:371.
39. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res* 2008;**36**:6882–92.
40. Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, et al. SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res* 2010;**38**:W487–96.
41. Ansari MZ, Yadav G, Gokhale RS, Mohanty D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res* 2004;**32**:W405–13.
42. Yadav G, Gokhale RS, Mohanty D. SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res* 2003;**31**: 3654–8.
43. Baranasic D, Zucko J, Diminic J, Gacesa R, Long PF, Cullum J, et al. Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J Ind Microbiol Biotechnol* 2014;**41**:461–7.
44. Khayatt BI, Overmars L, Siezen RJ, Francke C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS ONE* 2013;**8**:e62136.
45. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* 2011;**39**:W362–7.
46. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res* 2005;**33**:5799–808.
47. Prieto C, Garcia-Estrada C, Lorenzana D, Martin JF. NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* 2012;**28**:426–7.
48. Kamra P, Gokhale RS, Mohanty D. SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. *Nucleic Acids Res* 2005;**33**:W220–5.
49. Knudsen M, Sondergaard D, Tofting-Olesen C, Hansen FT, Brodersen DE, Pedersen CN. Computational discovery of specificity-conferring sites in non-ribosomal peptide synthetases. *Bioinformatics* 2015. doi: 10.1093/bioinformatics/btv600.
50. Hammami R, Zouhir A, Ben Hamida J, Fliss I. BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiol* 2007;**7**:89.
51. Hammami R, Zouhir A, Le Lay C, Ben Hamida J, Fliss I. BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol* 2010;**10**:22.
52. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res* 2013;**41**:D402–7.
53. Diminic J, Zucko J, Ruzic IT, Gacesa R, Hranueli D, Long PF, et al. Databases of the thiotemplate modular systems (CSDB) and their *in silico* recombinants (r-CSDB). *J Ind Microbiol Biotechnol* 2013;**40**:653–9.
54. Ichikawa N, Sasagawa M, Yamamoto M, Komaki H, Yoshida Y, Yamazaki S, et al. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 2013;**41**:D408–14.
55. Hadjithomas M, Chen IM, Chu K, Ratner A, Palaniappan K, Szeto E, et al. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* 2015;**6**:e00932.
56. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* 2015;**11**:625–31.
57. Starcevic A, Wolf K, Diminic J, Zucko J, Ruzic IT, Long PF, et al. Recombinatorial biosynthesis of polyketides. *J Ind Microbiol Biotechnol* 2012;**39**:503–11.
58. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2015;**44**(D1):D1214–9.
59. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* 2015;**43**:W612–20.
60. Kelly R, Kidd R. Editorial: ChemSpider – a tool for natural products research. *Nat Prod Rep* 2015;**32**:1163–4.
61. Nakamura Y, Afendi FM, Parvin AK, Ono N, Tanaka K, Hirai Morita A, et al. KNApSAcK Metabolite Activity Database for retrieving the relationships between metabolites and biological activities. *Plant Cell Physiol* 2014;**55**:e7.
62. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, et al. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* 2012;**53**:e1.
63. Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res* 2008;**36**:D326–31.
64. Flissi A, Dufresne Y, Michalik J, Tonon L, Janot S, Noe L, et al. Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic Acids Res* 2015;**44**(D1):D1113–8.
65. Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. In: Wheeler R, Spellmeyer D, editors. Annual reports in computational chemistry, Vol. 4. Washington, DC: American Chemical Society; 2008. p. 217–41.
66. Lucas X, Senger C, Erxleben A, Gruning BA, Doring K, Mosch J, et al. StreptomeDB: a resource for natural compounds isolated from *Streptomyces* species. *Nucleic Acids Res* 2013;**41**:D1130–6.
67. Klementz D, Doring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, et al. StreptomeDB 2.0 – an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res* 2016;**44**(D1):D509–14.
68. Mohimani H, Liu WT, Mylne JS, Poth AG, Colgrave ML, Tran D, et al. Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. *J Proteome Res* 2011;**10**:4505–12.
69. Ibrahim A, Yang L, Johnston C, Liu X, Ma B, Magarvey NA. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc Natl Acad Sci U S A* 2012;**109**:19196–201.
70. Yang L, Ibrahim A, Johnston CW, Skinnider MA, Ma B, Magarvey NA. Exploration of nonribosomal peptide families with an automated informatic search algorithm. *Chem Biol* 2015;**22**:1259–69.
71. Johnston CW, Connaty AD, Skinnider MA, Li Y, Grunwald A, Wyatt MA, et al. Informatic search strategies to discover analogues and variants of natural product archetypes. *J Ind Microbiol Biotechnol* 2015;doi:10.1007/s10295-015-1675-9.
72. Mohimani H, Liu WT, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *J Nat Prod* 2014;**77**:1902–9.
73. Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput Biol* 2014;**10**:e1003822.
74. Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, et al. Automated genome mining of ribosomal peptide natural products. *ACS Chem Biol* 2014;**9**:1545–51.
75. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
76. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**:e1002195.
77. Medema MH, Takano E, Breitling R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 2013;doi:10.1093/molbev/mst025.
78. Skinnider MA, Johnston CW, Zvanych R, Magarvey NA. Automated identification of depsipeptide natural products by an informatic search algorithm. *Chembiochem* 2015;**16**:223–7.
79. Blin K, Kazempour D, Wohlleben W, Weber T. Improved lanthipeptide detection and prediction for antiSMASH. *PLoS ONE* 2014;**9**:e89420.

80. Stachelhaus T, Mootz HD, Marahiel MA. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol* 1999;**6**:493–505.
81. Challis GL, Ravel J, Townsend CA. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* 2000;**7**:211–24.
82. Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J Mol Biol* 2007;**368**:1500–17.
83. Lee TV, Johnson RD, Arcus VL, Lott JS. Prediction of the substrate for nonribosomal peptide synthetase (NRPS) adenylation domains by virtual screening. *Proteins* 2015;**83**(11):2052–66.
84. Vijayan M, Chandrika SK, Vasudevan SE. PKSIIIexplorer: TSVM approach for predicting Type III polyketide synthase proteins. *Bioinformation* 2011;**6**:125–7.
85. Charlop-Powers Z, Owen JG, Reddy BV, Ternei MA, Brady SF. Chemical-biogeographic survey of secondary metabolism in soil. *Proc Natl Acad Sci U S A* 2014;**111**:3757–62.
86. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 2014;**42**:D1083–90.
87. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
88. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 2013;**41**:D456–63.
89. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**:D344–50.
90. Wu C, Choi YH, van Wezel GP. Metabolic profiling as a tool for prioritizing antimicrobial compounds. *J Ind Microbiol Biotechnol* 2015;doi:10.1007/s10295-015-1666-x.
91. Wu C, Kim HK, van Wezel GP, Choi YH. Metabolomics in the natural products field – a gateway to novel antibiotics. *Drug Discov Today Technol* 2015;**13**:11–7.
92. Kim HU, Kim TY, Lee SY. Metabolic flux analysis and metabolic engineering of microorganisms. *Mol Biosyst* 2008;**4**:113–20.
93. Lule I, Huys PJD, Van Mellaert L, Anne J, Bernaerts K, Van Impe J. Metabolic impact assessment for heterologous protein production in *Streptomyces lividans* based on genome-scale metabolic network modeling. *Math Biosci* 2013;**246**:113–21.
94. Huys PJD, Lule I, Vercammen D, Anne J, Van Impe JF, Bernaerts K. Genome-scale metabolic flux analysis of *Streptomyces lividans* growing on a complex medium. *J Biotechnol* 2012;**161**:1–13.
95. Kim M, Sang Yi J, Kim J, Kim JN, Kim MW, Kim BG. Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in *Streptomyces coelicolor* A3(2). *Biotechnol J* 2014;**9**:1185–94.
96. Huang D, Li S, Xia M, Wen J, Jia X. Genome-scale metabolic network guided engineering of Streptomyces tsukubaensis for FK506 production improvement. *Microb Cell Fact* 2013;**12**:52.
97. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 2010;**5**:93–121.
98. Alam MT, Medema MH, Takano E, Breitling R. Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism. *FEBS Lett* 2011;**585**:2389–94.
99. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010;**28**:977–82.
100. Pabinger S, Rader R, Agren R, Nielsen J, Trajanoski Z. MEMOSys: bioinformatics platform for genome-scale metabolic models. *BMC Syst Biol* 2011;**5**:20.
101. Swainston N, Smallbone K, Mendes P, Kell D, Paton N. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform* 2011;**8**:186.
102. Boele J, Olivier BG, Teusink B. FAME, the Flux Analysis and Modeling Environment. *BMC Syst Biol* 2012;**6**:8.
103. Liao YC, Tsai MH, Chen FC, Hsiung CA. GEMSiRV: a software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics* 2012;**28**:1752–8.
104. Latendresse M, Krummenacker M, Trupp M, Karp PD. Construction and completion of flux balance models from pathway databases. *Bioinformatics* 2012;**28**:388–96.
105. Feng X, Xu Y, Chen Y, Tang YJ. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC Syst Biol* 2012;**6**:94.
106. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol* 2013;**9**:e1002980.
107. Pitkanen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, et al. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol* 2014;**10**:e1003465.
108. Dias O, Rocha M, Ferreira EC, Rocha I. Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res* 2015;**43**:3899–910.
109. Hamilton JJ, Reed JL. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ Microbiol* 2014;**16**:49–59.
110. Gubbens J, Zhu H, Girard G, Song L, Florea BI, Aston P, et al. Natural product proteomining, a quantitative proteomics platform, allows rapid discovery of biosynthetic gene clusters for different classes of natural products. *Chem Biol* 2014;**21**:707–18.
111. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 2014;**10**:963–8.
112. Frasch HJ, Medema MH, Takano E, Breitling R. Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice. *Curr Opin Biotechnol* 2013;**24**:1144–50.
113. Medema MH, van Raaphorst R, Takano E, Breitling R. Computational tools for the synthetic design of biochemical pathways. *Nat Rev Microbiol* 2012;**10**:191–202.
114. Del Vecchio F, Petkovic H, Kendrew SG, Low L, Wilkinson B, Lill R, et al. Active-site residue, domain and module swaps in modular polyketide synthases. *J Ind Microbiol Biotechnol* 2003;**30**:489–94.
115. Menzella HG, Reid R, Carney JR, Chandran SS, Reisinger SJ, Patel KG, et al. Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. *Nat Biotechnol* 2005;**23**:1171–6.
116. Nguyen KT, Ritz D, Gu JQ, Alexander D, Chu M, Miao V, et al. Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc Natl Acad Sci U S A* 2006;**103**:17462–7.
117. Butz D, Schmiederer T, Hadatsch B, Wohlleben W, Weber T, Süßmuth RD. Module extension of a non-ribosomal peptide synthetase of the glycopeptide antibiotic balhimycin produced by *Amycolatopsis balhimycina*. *Chembiochem* 2008;**9**:1195–200.
118. Kapur S, Lowry B, Yuzawa S, Kenthirapalan S, Chen AY, Cane DE, et al. Reprogramming a module of the 6-deoxyerythronolide B synthase for iterative chain elongation. *Proc Natl Acad Sci U S A* 2012;**109**:4110–5.
119. McDaniel R, Thamchaipenet A, Gustafsson C, Fu H, Betlach M, Ashley G. Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A* 1999;**96**:1846–51.
120. Tanovic A, Samel SA, Essen LO, Marahiel MA. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science* 2008;**321**:659–63.
121. Marahiel MA. A structural model for multimodular NRPS assembly lines. *Nat Prod Rep* 2015;doi:10.1039/C5NP00082C.
122. Whicher JR, Dutta S, Hansen DA, Hale WA, Chemler JA, Dosey AM, et al. Structural rearrangements of a polyketide synthase module during its catalytic cycle. *Nature* 2014;**510**:560–4.
123. Dutta S, Whicher JR, Hansen DA, Hale WA, Chemler JA, Congdon GR, et al. Structure of a modular polyketide synthase. *Nature* 2014;**510**:512–7.
124. Kapur S, Chen AY, Cane DE, Khosla C. Molecular recognition between ketosynthase and acyl carrier protein domains of the 6-deoxyerythronolide B synthase. *Proc Natl Acad Sci U S A* 2010;**107**:22066–71.
125. Ye Z, Musiol EM, Weber T, Williams GJ. Reprogramming acyl carrier protein interactions of an acyl-CoA promiscuous *trans*-acyltransferase. *Chem Biol* 2014;**21**:636–46.
126. Tong Y, Charusanti P, Zhang L, Weber T, Lee SY. CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS Synth Biol* 2015;**4**:1020–9.
127. Cobb RE, Wang Y, Zhao H. High-efficiency multiplex genome editing of *Streptomyces* species using an engineered CRISPR/Cas system. *ACS Synth Biol* 2015;**4**:723–8.
128. Zeng H, Wen S, Xu W, He Z, Zhai G, Liu Y, et al. Highly efficient editing of the actinorhodin polyketide chain length factor gene in *Streptomyces coelicolor* M145 using CRISPR/Cas9-CodA(sm) combined system. *Appl Microbiol Biotechnol* 2015;**99**(24):10575–85.
129. Huang H, Zheng G, Jiang W, Hu H, Lu Y. One-step high-efficiency CRISPR/Cas9-mediated genome editing in *Streptomyces*. *Acta Biochim Biophys Sin (Shanghai)* 2015;**47**:231–43.